Stochastic optimization using automatic relevance determination prior model for Bayesian compressive sensing

Yong Huang*[a,b], James L Beck [b], Stephen Wu [b], Hui Li [a]
[a]School of Civil Engineering, Harbin Institute of Technology, 73 Huanghe Road, Harbin, People's Republic of China 150090.
[b]Division of Engineering and Applied Science, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125 USA

ABSTRACT

Compared with the conventional monitoring approach of separately sensing and then compressing the data, compressive sensing (CS) is a novel data acquisition framework whereby the compression is done during the sampling. If the original sensed signal would have been sufficiently sparse in terms of some orthogonal basis, the decompression can be done essentially perfectly up to some critical compression ratio. In structural health monitoring (SHM) systems for civil structures, novel data compression techniques such as CS are needed to reduce the cost of signal transfer and storage. In this article, Bayesian compressive sensing (BCS) is investigated for SHM signals. By explicitly quantifying the uncertainty in the signal reconstruction, the BCS technique exhibits an obvious benefit over the existing regularized norm-minimization CS. However, current BCS algorithms suffer from a robustness problem; sometimes the reconstruction errors are large. The source of the problem is that inversion of the compressed signal is a severely ill-posed problem that often leads to sub-optimal signal representations. To ensure the strong robustness of the signal reconstruction, even at a high compression ratio, an improved BCS algorithm is proposed which uses stochastic optimization for the automatic relevance determination approach to reconstructing the underlying signal. Numerical experiments are used as examples; the improved BCS algorithm demonstrates superior performance than state-of-the-art BCS reconstruction algorithms.

Keywords: Bayesian compressive sensing, Data compression, Structural health monitoring, Relevance vector machine, Automatic Relevance Determination, Robustness, Stochastic optimization, Simulated annealing

1. INTRODUCTION

A substantial number of sensors are required for structural health monitoring (SHM) systems due to the complexity and large scale of civil structures. Consequently, a large amount of data is usually produced by SHM systems. Therefore, data compression is necessary to reduce the cost and increase the efficiency of signal transfer and storage. Data compression [1-2] for SHM systems has attracted much interest in recent years, especially for wireless monitoring systems [3], since data compression techniques can provide a way to improve the power efficiency and minimize bandwidth during the transmission of structural response time-histories from wireless sensors [4]. All of these data compression methods belong to a conventional framework for sampling signals that follow Shannon's celebrated theorem: the sampling rate must be at least twice the maximum frequency present in the signal.

Compressive sensing (CS) [5-6] is a novel sampling technique that goes against the common wisdom in data acquisition. It asserts that if certain signals are sparse in some orthogonal basis, one can reconstruct these signals from far fewer measurements than what is usually considered necessary based on Nyquist-Shannon sampling theory. This new technique may come to underlie procedures for sampling and compressing data simultaneously, therefore increasing the efficiency of data transfer and storage.

In this article, we utilize a Bayesian regression model for the CS problem and investigate the robustness of the signal reconstruction. An improved Bayesian CS (BCS) algorithm is developed to reduce the likelihood of sub-optimal solutions of the optimization problem during the reconstruction process and so to produce smaller reconstruction errors.

*yonghuan@caltech.edu; phone 1 626-395-3491;

Health Monitoring of Structural and Biological Systems 2012, edited by Tribikram Kundu, Proc. of SPIE Vol. 8348, 834837 · © 2012 SPIE · CCC code: 0277-786X/12/$18 · doi: 10.1117/12.921257

2. BAYESIAN COMPRESSIVE SENSING

2.1 Bayesian linear regression model for compressive sensing

Consider a discrete-time signal $\mathbf{x} = [x(1), \cdots x(N)]^T$ in $\mathbb{R}^N$ represented in terms of a set of orthogonal basis vector as

$$\mathbf{x} = \sum_{n=1}^{N} w_n \boldsymbol{\Psi}_n \text{ or } \mathbf{x} = \boldsymbol{\Psi}\mathbf{w} \quad (1)$$

where $\boldsymbol{\Psi} = [\boldsymbol{\Psi}_1, \cdots, \boldsymbol{\Psi}_N]$ is the $N \times N$ basis matrix with the orthonormal basis of $N \times 1$ vectors $\{\boldsymbol{\Psi}_n\}_{n=1}^N$ as columns; $\mathbf{w}$ is the sparse coefficients or weight vector, i.e., it is known that most of its components are zero or very small (with minimal impact on the signal) but not which ones.

In the framework of CS, one infers the coefficients w_n of interest from compressed data instead of directly sampling the signal $\mathbf{x}$. The data vector $\mathbf{y}$ from the compressive sensor is composed of K individual measurements obtained by linearly projecting the signal $\mathbf{x}$, which is sparse in the basis $\{\Psi_n\}$, by using a chosen and fixed random projection matrix $\boldsymbol{\Phi}$ (each element is i.i.d $\mathbf{N}(0,1)$):

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{x} + \mathbf{r} = \boldsymbol{\Theta}\mathbf{w} + \mathbf{e} = \sum_{n=1}^{N} w_n \boldsymbol{\Theta}_n + \mathbf{e} \quad (2)$$

where $\boldsymbol{\Theta} = \boldsymbol{\Phi}\boldsymbol{\Psi}$ is known and $\mathbf{e}$ represents the unknown prediction error due to the signal model for specified $\mathbf{w}$ plus any measurement error $\mathbf{r}$; $\mathbf{e}$ is modeled as a zero-mean Gaussian vector with covariance matrix $\sigma^2\mathbf{I}_K$. For the purpose of data compression, $\boldsymbol{\Theta}$ is a $K \times N$ matrix with $K \ll N$, which leads to an ill-posed inversion problem for finding the weights $\mathbf{w}$ and hence the signal $\mathbf{x}$ in $\mathbb{R}^N$ from data $\mathbf{y}$ in $\mathbb{R}^K$.

By exploiting the sparsity of the representation of $\mathbf{x}$ in basis $\{\Psi_n\}_{n=1}^N$, the ill-posed problem can be posed as a convex optimization problem to estimate $\mathbf{w}$ as follow:

$$\widetilde{\mathbf{w}} = \arg\ \min\{\|\mathbf{y} - \boldsymbol{\Theta}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1\} \quad (3)$$

where the penalty parameter λ scales the regularization term to penalize large weight values. As a result, the optimization problem in (3) represents the trade-off between how well the data is fitted (first term) and how sparse the signal is (second term). Some norm minimization algorithms have been proposed to solve the CS reconstruction problem as formulated in (3) [7-9]. The choice of the 1-norm regularization in (3) is very important because it induces sparsity in the $\widetilde{\mathbf{w}}$ while still giving a convex optimization problem.

The ill-posed data inverse problem can also be tackled using a Bayesian perspective, which has certain distinct advantages compared to previously published CS inversion algorithms; for example, in addition to providing a sparse solution to estimate the underlying signal, it automatically estimates penalty parameters and it provides a measure of the uncertainty for the reconstructed signal.

Ji et al. [10] adopt the ideas of sparse Bayesian learning proposed in [11-12] for regression to solve the CS reconstruction problem. The basic idea of the Bayesian approach is to apply Bayes's Theorem to find the posterior probability density function (PDF) for the signal weights based on the linearly projected data by multiplying a prior PDF for the weights with a likelihood function for the data and then normalizing the product. For compressive sensing, because of the probability model for prediction error $\mathbf{e}$ in (2), one gets a Gaussian likelihood function:

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{\frac{K}{2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \boldsymbol{\Theta}\mathbf{w}\|_2^2\right) \quad (4)$$

The likelihood here measures how well the signal model for specified parameters $\mathbf{w}$ and σ^2 predicts the observed CS measurements $\mathbf{y}$, and it corresponds to the first term of (3) in the deterministic CS formulation.

To induce sparsity, sparse Bayesian learning introduces the automatic relevance determination (ARD) prior, which is a multiple of N independent Gaussian priors, one for each w_n with a corresponding hyperparameter α_n:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{n=1}^{N}\left[(2\pi)^{-1/2}{\alpha_n}^{1/2} exp\left\{-\frac{1}{2}\alpha_n w_n^2\right\}\right] \quad (5)$$

Rather than finding the MAP (most probable) value of $\mathbf{w}$ based on data $\mathbf{y}$, a full Bayesian treatment is used for $\mathbf{w}$ by "marginalizing" $\mathbf{w}$ out and the MAP value of the α_n's is instead sought (usually with a uniform prior on the α_n's).

2.2 Bayesian compressive sensing reconstruction

Given the CS measurements $\mathbf{y}$ and the prior, the posterior distribution $p(\mathbf{w}|\mathbf{y},\boldsymbol{\alpha},\sigma^2)$ over the weights is obtained based on Bayes' theorem

$$p(\mathbf{w}|\mathbf{y},\boldsymbol{\alpha},\sigma^2) = p(\mathbf{y}|\mathbf{w},\sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})/p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2) \tag{6}$$

where $p(\mathbf{w}|\boldsymbol{\alpha})$=prior PDF of $\mathbf{w}$ in (5); $p(\mathbf{y}|\mathbf{w},\sigma^2)$=likelihood in (4); and $p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)$ is called the evidence of the signal model class $\mathcal{M}(\boldsymbol{\alpha},\sigma^2)$ which also serves as a normalizing constant for the posterior distribution.

Since both prior and likelihood for $\mathbf{w}$ are Gaussian and the likelihood mean $\boldsymbol{\Theta}\mathbf{w}$ is linear in $\mathbf{w}$, the posterior PDF can be expressed analytically as a multivariate Gaussian distribution $\mathrm{p}(\mathbf{w}|\mathbf{y},\boldsymbol{\alpha},\sigma^2) = N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ with mean and covariance [11][12]:

$$\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Theta}^{\mathbf{T}}\mathbf{y} \tag{7}$$

$$\boldsymbol{\Sigma} = (\sigma^{-2}\boldsymbol{\Theta}^T\boldsymbol{\Theta} + \mathbf{A})^{-1} \tag{8}$$

In the next step, Bayesian model class assessment is used to select the most plausible hyperparameters $\boldsymbol{\alpha}$ and σ^2. If the problem is globally identifiable [14], meaning here that the evidence $p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)$ has a single pronounced global maximum with respected to $\boldsymbol{\alpha}$ and σ^2, the reconstruction can be done using the most probable model class $\mathcal{M}(\hat{\boldsymbol{\alpha}},\hat{\sigma}^2)$ based on measurements $\mathbf{y}$, that is, by finding $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}^2$ that maximize $p(\boldsymbol{\alpha},\sigma^2|y) \propto p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)p(\boldsymbol{\alpha},\sigma^2)$. If a uniform prior on $\boldsymbol{\alpha}$ and σ^2 is considered, it is equivalent to the maximization of the evidence $p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)$, which has the following form [13]:

$$p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2) = \int p(\mathbf{y}|\mathbf{w},\sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w}$$

$$= (2\pi)^{-K/2}|\,\sigma^2\mathbf{I} + \boldsymbol{\Theta}\mathbf{A}^{-1}\boldsymbol{\Theta}^{\mathrm{T}}|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{y}^T(\,\sigma^2\mathbf{I} + \boldsymbol{\Theta}\mathbf{A}^{-1}\boldsymbol{\Theta}^T)^{-1}\mathbf{y}\right\} \tag{9}$$

A "Bottom-up" or "Fast" Algorithm [13] has been proposed for the original sparse Bayesian learning algorithm to find $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}^2$. The algorithm starts with no terms in (1) and adds relevant ones to the signal model as the iterations proceed. This method can significantly reduce the reconstruction time and the chance of having ill-conditioning problems during inversion of the Hessian matrix. In this algorithm, the log evidence $\log p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)$ is expressed as

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\alpha},\sigma^2) = \log p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)\\
&=-\frac{1}{2}[K\log 2\pi + \log|\mathbf{C}| + y^T\mathbf{C}^{-1}y]\\
&=-\frac{1}{2}\left[K\log 2\pi + \log|\mathbf{C}_{-n}| + y^T\mathbf{C}_{-n}^{-1}y - \log\alpha_n + \log(\alpha_n + \Theta_n^T\mathbf{C}_{-n}^{-1}\Theta_n) - \frac{(\Theta_n^T\mathbf{C}_{-n}^{-1}y)^2}{\alpha_n+\Theta_n^T\mathbf{C}_{-n}^{-1}\Theta_n}\right]\\
&=\mathcal{L}(\boldsymbol{\alpha}_{-n},\sigma^2)+\frac{1}{2}\left[\log\alpha_n + \log(\alpha_n + \Theta_n^T\mathbf{C}_{-n}^{-1}\Theta_n) - \frac{(\Theta_n^T\mathbf{C}_{-n}^{-1}y)^2}{\alpha_n+\Theta_n^T\mathbf{C}_{-n}^{-1}\Theta_n}\right]\\
&=\mathcal{L}(\boldsymbol{\alpha}_{-n},\sigma^2)+\frac{1}{2}\left[\log\alpha_n + \log(\alpha_n + S_n) - \frac{Q_n^2}{\alpha_n+S_n}\right]
\end{aligned} \tag{10}$$

where $\mathbf{C} = \sigma^2 I + \boldsymbol{\Theta}\mathbf{A}^{-1}\boldsymbol{\Theta}^T$ and $\mathbf{C}_{-n}$=covariance matrix $\mathbf{C}$ with the components of n removed, and therefore $\mathcal{L}(\boldsymbol{\alpha}_{-n},\sigma^2)$ does not depend on α_n. Besides, for simplification of forthcoming expressions, we have defined the 'sparsity factor' $S_n$and 'quality factor' Q_n by:

$$S_n = \Theta_n^T\mathbf{C}_{-n}^{-1}\Theta_n \tag{11}$$

$$Q_n = \Theta_n^T\mathbf{C}_{-n}^{-1}y. \tag{12}$$

Setting derivatives of (10) with respect to α_n to zero leads to

$$\hat{\alpha}_n = \begin{cases}\infty, & \text{if } Q_n^2 \le S_n\\ \frac{S_n^2}{Q_n^2-S_n}, & \text{if } Q_n^2 > S_n\end{cases} \tag{13}$$

This algorithm enables an efficient sequential optimization by updating one candidate basis term at each iteration to monotonically increase the evidence. Finally, only the components that have finite α_n are used in determining the signal model [13].

2.3 Robustness of BCS reconstruction

In the BCS reconstruction, the number of the measurements **y** has an important influence on the robustness of signal reconstruction. However, in order to compress data more effectively, the number of the measurements must be reduced to be much smaller than the number of degrees of freedom of the original signal. As a result, inversion of the signal becomes a severely ill-posed problem that leads to sub-optimal signal representations. This occurs because there are a large number of local maxima that trap the optimization and significantly reduce the robustness of the iterative scheme.

Figure 1 demonstrates the reconstruction error and log evidence as a function of size of the final reconstructed signal models (the number of non-zero terms in (1)) which are reconstructed from 1000 different random measurement samples of size K (by changing the projection matrix $\mathbf{\Phi}$) using the same original signal (a constant noise variance $\sigma^2 = \mathrm{var}[\mathbf{y}] \times 0.1$ is chosen for all trials).

In this figure, we consider signals of length N = 512, each containing 20 non-zeros spikes created by randomly choosing 20 discrete times; the non-zero spikes of the signals are drawn from a zero-mean unit variance Gaussian distribution. We tried to set the number of measurements K so that around one half of the trials gave a correct reconstruction and this gave $\mathrm{K} = 70$ (corresponding to a compression ratio of 7.3). The reconstruction error in Figure 1 is defined as $\|\hat{\mathrm{x}} - \mathrm{x}\|_2 / \|\mathrm{x}\|_2$, where $\hat{\mathrm{x}}$ denotes the reconstructed signal.

The figure shows that many of the optimization runs produced the correct signal size of 20, and it was found that all of these reconstructions are quite close to a global maximum of the log evidence. A certain number of the runs give local maxima of the evidence that correspond to larger amounts of non-zero signal components. Figure 1 illustrates that the reconstruction process has poor robustness.

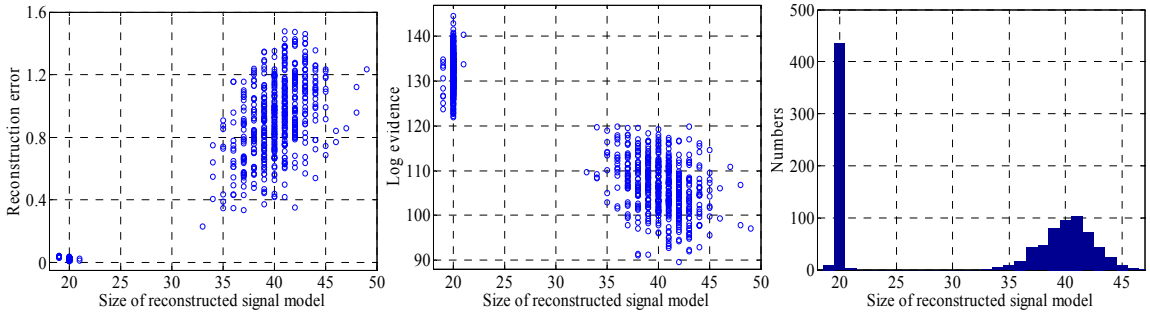

Figure 1. Reconstruction error, log evidence and numbers of cases for N=512, T=20 as a function of size of the reconstructed signal models for the noise-free case.

3. PROPOSED METHOD

3.1 Modified BCS using a stochastic optimization approach

In the Bottom-up Algorithm of sparse Bayesian learning, adding, deleting or re-estimation for a basis vector for each iteration of the optimization over the individual hyperparameters is based on whatever gives the maximum evidence increase. This is a deterministic optimization procedure. In this fast algorithm, it is found that the value of the evidence change $\Delta\mathcal{L}$ is sensitive to the details of the specific choice of projection matrix and the corresponding measurements when the number of measurements is smaller than a specific threshold. The fewer the number of measurements, the more sensitive the evidence is to the details. This is because the evidence function has a lot of local maxima when the ratio K/N falls below a specific threshold. Thus, there exists no one particular action (add, delete or re-estimate) that leads to an exceptionally large evidence increase $\Delta\mathcal{L}$ among the set of $\Delta\mathcal{L}$ for all possible actions. This means that each time an action is chosen to maximize $\Delta\mathcal{L}$ in each iterative step of BCS, it is very sensitive to any slight changes in the problem

setting, such as the choice of projection matrix. Therefore, this algorithm shows a low robustness for finding the global maximum of the evidence over the hyperparameters that produces the optimal signal reconstruction.

This consideration motivates trying a stochastic method for optimization based on the ARD prior model, by introducing uncertainty into the optimization direction. One of the roles of the injected uncertainty in the proposed stochastic optimization process is to allow for a broader search of potential unexplored areas which may contain the global maximum. Therefore, this approach significantly increases the probability of finding this global maximum, even when the evidence function has many local maxima.

We set the probability of accepting the action of adding, deleting or re-estimation of the candidate n^{th} basis vector as

$$p_n = \Delta\mathcal{L}_n / \max(\Delta\mathcal{L}_n) \tag{14}$$

where $\Delta\mathcal{L}_n$ is the change of evidence produced by the n^{th} basis vector being added, deleted or re-estimated and $\max(\Delta\mathcal{L}_n)$ is the largest increase of the log evidence in a specific step. Thus, if the evidence change $\Delta\mathcal{L}_n$is the largest among all basis vectors (p_n =1), then definitely accept the action of this candidate basis vector. If the evidence change $\Delta\mathcal{L}_n$ is not the largest (p_n <1), then accept the action of the candidate n^{th}basis function with probability p_n. We can accomplish this process by generating an independent value u_n from the uniform distribution $U(0,1)$. If $p_n \geq u_n$, the corresponding action of the n^{th} basis function is implemented, otherwise it is skipped. Finally, we use the same termination criterion as the Bottom-up Algorithm, which judges convergence by $\max(\Delta\mathcal{L}_n)$ being sufficiently small.

3.2 An analogue to simulated annealing

Reducing the uncertainty of the reconstructed results is an important goal for a robust BCS reconstruction algorithm. Motivated by the idea of incorporating simulated annealing in many MCMC algorithms [15-16], which can adaptively control the Kullback-Leibler information between the posterior and prior PDFs, a step-wise optimization approach is introduced to utilize a special structure of the evidence function in our problem.

In the fast BCS algorithm, the initial guess of the σ^2 for the iterative scheme may affect the algorithm significantly due to the underdetermined nature of the inverse problem. The variance σ^2 has significant influence on the trade-off between how well the reconstructed signal model fits the data and how sparse it is. It is found the evidence function in our case tends to have more significant local maxima for a fixed value of σ^2 as we decrease σ^2. Hence, the set of evidence functions for a given set of fixed σ^2 exhibits a similar structure to the intermediate PDFs under MCMC with simulated annealing. First, we initialize the algorithm with a large σ^2 and optimize the intermediate evidence function (corresponding to the fixed σ^2) to obtain a set of intermediate optimal hyperparameters. Because the intermediate evidence function tends to be smoother, there is a higher probability for the intermediate optimal hyperparameters $[\alpha_1, \alpha_2, \ldots \alpha_{n,} \ldots, \alpha_N]$ to be near the global maximum. However, then the data fitting ability is poor and so extra optimization is needed to ensure better fitting of the data. This can be achieved by the procedure in Section 3.1 as it allows a non-zero probability to accept candidates with higher cost that helps the algorithm to escape being trapped in local maxima.

At each intermediate step, once the intermediate optimum hyperparameters are found, the updated σ^2 for the next optimization step is estimated from [10-11]:

$$(\sigma^2)^{[j+1]} = \frac{\left\|\mathbf{y}-\boldsymbol{\Theta}\boldsymbol{\mu}^{[j]}\right\|^2}{K-\sum_{n=1}^{N'}\left(1-\alpha_n^{[j]}\boldsymbol{\Sigma}_{nn}^{[j]}\right)} \qquad j = 1,2,3 \ldots J-1 \tag{15}$$

where N_j is the current size of the signal model; and $\boldsymbol{\mu}^{[j]}$ and $\boldsymbol{\Sigma}^{[j]}$ are the mean and covariance of the multivariate Gaussian distribution over the reconstructed sparse signal w from the j^{th} iteration of the optimization, calculated by (7) and (8), respectively. The updated σ^2 will tend to get smaller until eventually convergence to the optimal value of the evidence function occurs.

The purpose of this step-wise algorithm is to avoid optimization of the hyperparameters directly using the evidence function over the full high-dimensional hyperparameter space; instead, a series of optimizations with smaller information gain is performed that leads to a smoother intermediate evidence function for each value of σ^2, so that the algorithm gradually converges to the global maximum of the evidence. Although the geometrical shape of the target evidence function is very complex and a large number of local maxima around the initial points can trap the iterative optimization, the initial optimization stage with large σ^2 involves a smoother evidence function and the change of function shape

between two adjacent optimization stages can be made small. This small change makes it possible to dilute the effect of local maxima of the evidence function and perform the optimization more robustly.

Finally, the outer loop of updating σ^2 converges to the point with a maximum of the evidence function with respect to σ^2. The whole procedure is terminated when the changes of σ^2 are sufficiently small; $e.g.\ (\sigma^2)^{[j+1]} - (\sigma^2)^{[j]} < 10^{-6}$.

3.3 BCS reconstruction algorithm using stochastic optimization based on ARD prior model

We can combine the previous two ideas to produce a BCS reconstruction method using stochastic optimization based on the ARD prior model. The outer loop updates the prediction-error variance σ^2, and the inner loop is the stochastic optimization procedure. The procedure is summarized below in Algorithm 1.

Algorithm 1.BCS-SO: Improved BCS with stochastic optimization

1. Inputs: $\mathbf{\Theta}, \mathbf{y}$; Outputs: mean and covariance of $\mathbf{w}$
2. Initialize σ^2 as a large value (*e.g.*$\gamma \times \mathrm{var(y)},\ \gamma \geq 20$)
3. **While** convergence criterion not met
4. Calculate p_n in (14) for every basis vector Θ_n and update Θ_n if $p_n \geq u_n$, where u_n is sampled from $U(0,1)$, otherwise go to step 8. Calculate S_n and Q_n from (11) and (12).
5. **If** $Q_n^2 - S_n > 0$ and $\alpha_n = \infty$, add Θ_n and update α_n using (13)
6. **If** $Q_n^2 - S_n > 0$ and $\alpha_n < \infty$, re-estimate $\alpha_n$using (13)
7. **If** $Q_n^2 - S_n \leq 0$ and $\alpha_n < \infty$, delete Θ_n and set $\alpha_n = \infty$
8. **End if**
9. Update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (7) and (8)
10. **End while**, the intermediate optimal hyperparameters are obtained
11. Update σ^2 using (15)
12. Set the obtained intermediate optimal hyperparameters as initial information for the next inner loop and repeat the above procedure (steps 3 to 10) until the σ^2 updating converges.

4. EXAMPLE RESULTS

4.1 Synthetic sparse spike signal

We denote the Bottom-up Algorithm in [13] as BCS-B and our modified algorithm (Algorithm 1) in Section 3.3 as BCS-SO. As a comparison, we also give the performance of another improved algorithm. For improving robustness to the parameter setting for the prediction-error variance σ^2, Ji et.al [17] integrated out the uncertainty in σ^2 from the model instead of estimating the optimal σ^2. The corresponding improved algorithm is denoted as BCB-IOE.

We consider signals which are the same as in Figure 1. They are zero-mean unit variance Gaussian spikes (Figure 2 (a)). Figure 2 (b) demonstrates reconstruction results for the signals using the original BCS algorithm while Figure 2 (c) shows the results for the BCS-SO algorithm when K=70. For BCS-SO, we set the parameter $\gamma = 50$. Because of insufficient number of measurements, sub-optimal signal representations are obtained for BCS (Figure 2 (b)). However, BCS-SO produces almost perfect reconstructions (Figure 2 (c)) with a compression ratio of 7.3. The error-bars (defined as $\pm$ one standard deviation) are also shown in the results, which is the uncertainty estimates of the reconstructed coefficients.

The average performance is investigated with different numbers of measurements K. In these examples, we fix the signal with length $\mathrm{N} = 512$ and the number of non-zero coefficients $\mathrm{T} = 20$, and vary K from 40 to 120. The $T = 20$ non-zero spikes of the signals are drawn from a zero-mean unit variance Gaussian distribution. A Gaussian random projection matrix is constructed for each experiment; the associated reconstruction errors are calculated as $\|\hat{\mathrm{x}} - \mathrm{x}\|_2^2 / \|\mathrm{x}\|_2^2$, where $\hat{\mathrm{x}}$

and x are the reconstructed and original signal vectors, respectively. Because of the randomness of the projected CS measurements in this set of experiments, we execute the experiment 100 times and report the average performance.

Firstly, the effect of parameter γ (step 2 of Algorithm 1) on the performance of BCS-SO is studied in Figure 3. It is seen that larger γ gives better reconstruction results. This is due to less local maxima occurring with larger γ, which increases the reconstruction robustness. However, the improvement diminishes with increased γ. In the following examples, we set γ=50.

To investigate how the number of measurements K affects the reconstruction performance, we compared the algorithms of BCB-SO ($\gamma = 50$), BCS-B and BCS-IOE in Figure 4. Regarding the performance comparison, the superiority of BCB-SO is demonstrated by its much lower reconstruction errors than the others. Notice that the critical value of K above which good reconstruction performance occurs is about 80 for the new BCS-SO algorithm, and about 90 for the other two algorithms.

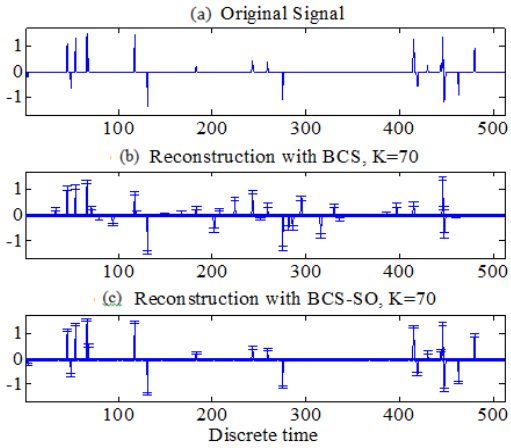

Figure 2. Original and reconstructed signals of length N=512.

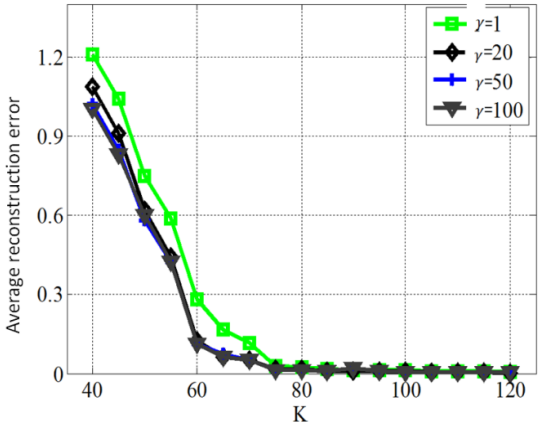

Figure 3. Comparison of performance using different values of parameter γ for BCS-SO.

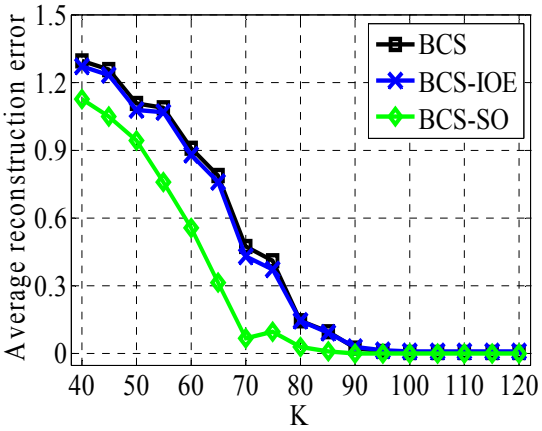

Figure.4. Reconstruction errors of BCS, BCB-IOE and BCB-SO as a function of increasing number of measurements K.

4.2 Application to bridge accelerometer data

Tianjin Yonghe Bridge (Figure 5) is one of the earliest cable-stayed bridges constructed in the mainland of China. It has a total length of 512 m, comprising a main span of 260m and two side spans of 125m. A sophisticated long-term structural health monitoring system was designed and implemented on this bridge by the Center of Structural Health Monitoring and Control (SMC) of the Harbin Institute of Technology during its rehabilitation. The system includes optical fiber Bragg-grating (FBG) strain sensors, accelerometers, electromagnetic sensors, GPSs, anemometer and temperature sensors. Fourteen uniaxial accelerometers and one biaxial accelerometer were permanently installed on the deck of the main span and two side spans, and on one tower top. Signals from one accelerometer which is installed on the deck of the main span are employed here.

Figure 5. Photo of the Yonghe Bridge.

Figure 6 shows the acceleration response time history from this accelerometer. The signal has 51200 samples at a sample frequency of 100 Hz. The acceleration data is divided into 100 segments of N=512, and each segment is decomposed by the dB1 wavelet basis with six resolution levels:

$$\mathbf{x} = \mathbf{\Psi w} \tag{16}$$

where $\mathbf{\Psi}$ is the dB1 wavelet basis matrix and $\mathbf{w}$ is the vector of wavelet coefficients. For convenience, the wavelet coefficients of the whole acceleration data are shown in Figure 7(a), which shows that only a small number (m) of the wavelet coefficients in $\mathbf{w}$ are significant, and the other ($N - m$) coefficients are small. After using soft threshold de-noising [18], the wavelet coefficients are sparse in the dB1 wavelet basis domain because only $m = 18400$ wavelet coefficients of the original 51200 coefficients are nonzero, as shown in Figure 7(b).

Therefore, the original acceleration can be expressed as:

$$\mathbf{x} = \mathbf{\Psi w}_d + \mathbf{\Psi n}_e = \mathbf{x}_d + \mathbf{\Psi n}_e \quad (17)$$

where $\mathbf{w}_d$ is the de-noised wavelet coefficients vector representing the original wavelet coefficients $\mathbf{w}$ with the smallest $N - m$ coefficients set to zero and $\mathbf{n}_e$ is the removed "noise" in the wavelet space. Also, $\mathbf{x}_d = \mathbf{\Psi w}_d$ denotes the de-noised acceleration data. Then the measurement vector can be expressed as:

$$\mathbf{y} = \mathbf{\Phi x} = \mathbf{\Theta w} = \mathbf{\Theta w}_d + \mathbf{\Theta n}_e = \mathbf{\Theta w}_d + \mathbf{e} \quad (18)$$

where $\mathbf{\Theta} = \mathbf{\Phi\Psi}$ is the projection matrix. In the reconstruction process, we model noise $\mathbf{e}$ as a zero-mean Gaussian vector with covariance matrix $\sigma^2\mathbf{I}_K$.

The discrete-time signal $\mathbf{x}$ in (17), and shown in Figure 6, is used to investigate the application of BCS to vibration signals used in SHM. In practice, one could acquire data already in a CS compressed form from a special sensor; for example, using modern wireless accelerometers where the CS algorithm is integrated with the ADC into the sensor itself. Then BCS reconstruction could be directly applied to the analog data from the sensors.

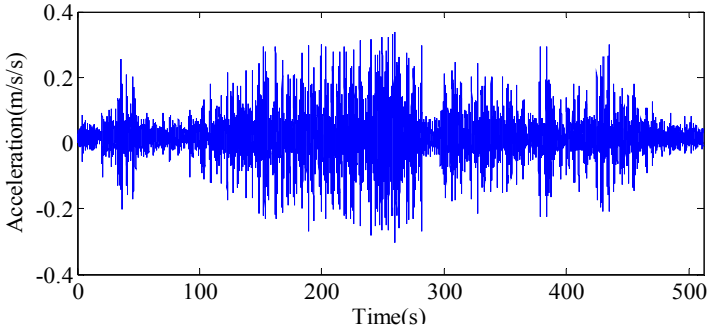

Figure 6. Acceleration response of bridge deck measured by accelerometer.

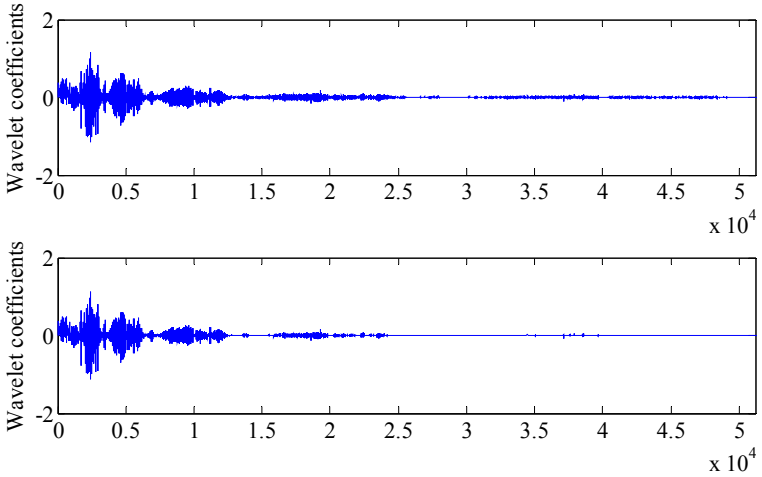

Figure 7. Wavelet coefficients of the acceleration data using the db1 wavelet basis: (a) original wavelet coefficients; (b) de-noised wavelet coefficients.

We observe from Figure 7 that the original wavelet coefficients $\mathbf{w}$ are contaminated with significant noise. Thus we employ the corresponding de-noised acceleration $\mathbf{x}_d = \mathbf{\Psi w}_d$ to obtain measurements $\mathbf{y}_d$ to investigate the performance of reconstruction algorithms for a case with lower noise, even though it is not practical in real CS applications. In this case, measurement vector $\mathbf{y}_d$ is expressed as

$$\mathbf{y}_d = \mathbf{\Phi x}_d = \mathbf{\Theta w}_d \quad (19)$$

We can obtain optimal reconstructed coefficients $\hat{\mathbf{w}}_d$ and $\hat{\mathbf{w}}$ from measurements $\mathbf{y}_d$ and $\mathbf{y}$, respectively, using the Bayesian reconstruction algorithm. The corresponding reconstructed accelerations $\hat{\mathbf{x}}_d$ and $\hat{\mathbf{x}}$ are then obtained by a wavelet transform using the reconstructed wavelet coefficients $\hat{\mathbf{w}}_d$ and $\hat{\mathbf{w}}$, respectively. The CS reconstruction errors

recovered from the measurement $\mathbf{y}$ and $\mathbf{y_d}$ are calculated by: $R_d = \|\mathbf{x_d} - \hat{\mathbf{x}}_{\boldsymbol{d}}\|_2 / \|\mathbf{x_d}\|_2$ and $R = \|\mathbf{x_d} - \hat{\mathbf{x}}\|_2 / \|\mathbf{x_d}\|_2$, respectively.

Also, we define a *sparsity rate index SR* to compare the sparsity in the reconstructed coefficients $\hat{\mathbf{w}}$ and original de-noised coefficients $\mathbf{w_d}$:

$$SR = \hat{s}/s_d \tag{20}$$

where $\hat{s}$ is the total number of significant coefficients which are larger than the threshold of 10^{-4} in vector $\hat{w}$, representing the sparsity of the reconstructed acceleration signal $\hat{x}$ with respect to the wavelet basis $\{\Psi_n\}$. s_d is the sparsity of the original de-noised wavelet coefficients vector w_d. If we obtain optimal reconstructed coefficient vector $\hat{\mathbf{w}}_d$, then the sparsity rate index is calculated as $SR_d = \hat{s}_d/s_d$, where $\hat{s}_d$ is the sparsity of the reconstructed acceleration signal $\hat{\mathbf{x}}_d$ with respect to the wavelet basis $\{\boldsymbol{\Psi}_n\}$.

Bao et.al [18] investigated the norm-minimization algorithm for CS reconstruction by using real acceleration data collected from a SHM system. In this study, BP algorithm [9] is also performed together with the BCS, BCS-IOE and BCS-SO algorithms to make a comparison of sparsity and reconstruction errors as shown in Figures 8 and 9. The algorithm BP is one of the norm minimization algorithms and so does not quantify the uncertainty in its reconstructed signals. We used the $l1$-magic package available online at http://www-stat.stanford.edu/~candes/l1magic/. In Figures 8 and 9, each evaluated point in the curves is computed based on the average of results of the 100 time segments and a different random projection matrix is chosen for each reconstruction. We define the reconstructions corresponding to measurements $\mathbf{y_d}$ and $\mathbf{y}$ as Case 1 and Case 2, respectively. Case 2 is the real case which should be tackled in structural health monitoring.

From the comparison of the results for the sparsity rate index in both cases, as shown in Figure 8, the BP algorithm tends to produce high under-sparsity ($SR > 1$). In contrast, the reconstructed coefficients vectors of the three BCS algorithms have a little over-sparsity as shown in Figure 8. This is beneficial for the central feature of BCS approaches, that the effective dimensionality of the signal model (equivalent to the number of retained coefficients) is determined automatically as part of the fully Bayesian inference procedure. It is a powerful advantage of such Bayesian approaches that they encourage sparsity of representation.

From the observation of the results of the reconstruction error R_d as shown in Figure 9, the proposed BCS-SO algorithm outperforms all other methods except for the first five chosen K values, for which it provides the best performance after BP, but BP produces worse performance than the other methods for the rest of the signals. For Case 2, it is seen that the BP performs better than all other Bayesian CS methods for the case of measurements $\mathbf{y}$, though the average reconstruction errors of BP and BCS-SO are close. Despite this fact, it is noted that the proposed BCS-SO method provides the best overall performance among all methods considering both the sparsity and uncertainty quantification of the results.

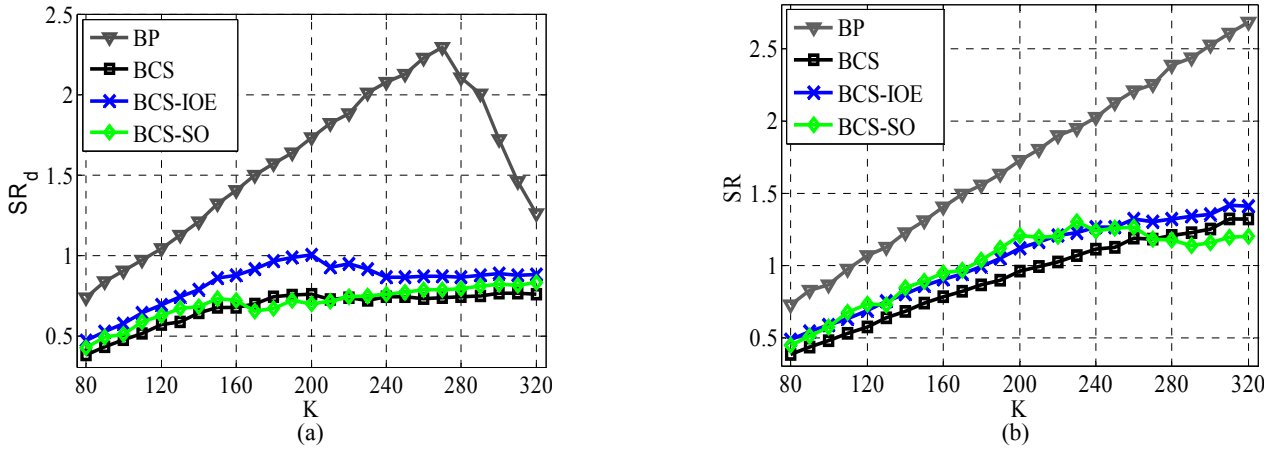

Figure 8. Sparsity ratio of four BCS algorithms as a function of K using real SHM acceleration data: (a). Case 1; (b). Case 2.

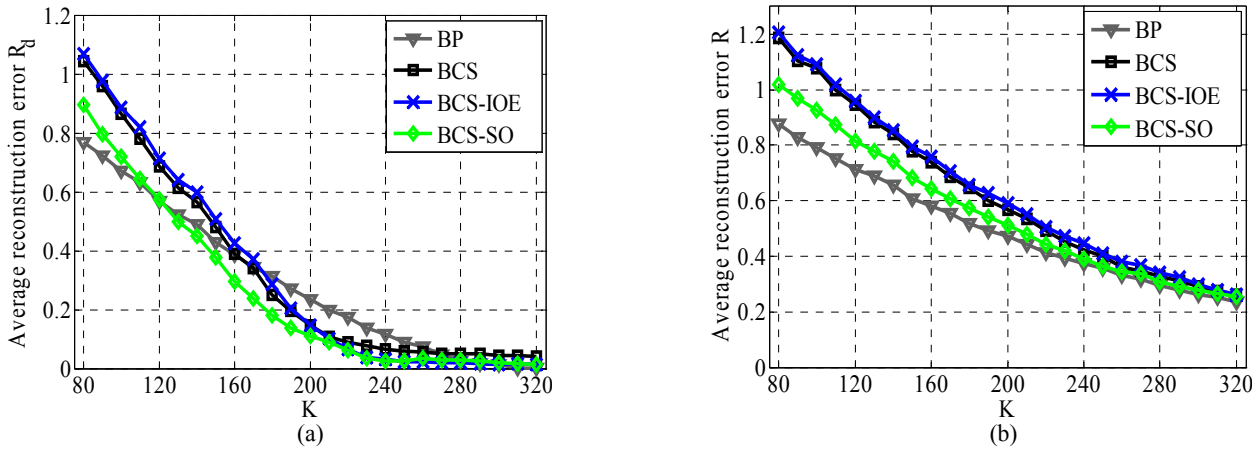

Figure 9. Reconstruction errors of four algorithms as a function of K using real SHM acceleration data: (a). Case 1; (b). Case 2.

5. CONCLUSION

In this paper, the improvement and application of the BCS technique for SHM signal is studied. We tackled the CS reconstruction problem using a Bayesian framework to estimate the sparse signal coefficients. We showed that when the number of measurements is much smaller than the length of the discrete-time signal, BCS reconstruction lacks robustness.

Based on these studies, an improved method which uses stochastic optimization and annealing with the ARD prior model is developed to reduce the chance of suboptimal signal representations. Both synthetic and real structural response signals are employed to validate the developed methods. It is demonstrated that overall the proposed BCS algorithm has a better performance than state-of-the-art BCS algorithms.

REFERENCES

[1] Xu, N., Rangwala, S., Chintalapudi, K., Ganesan, D., Broad, A., Govindan, R. and Estrin, D., "A wireless sensor network for structural monitoring," Proceedings of the ACM Conference on Embedded Networked Sensor Systems, Baltimore, MD, USA (2004).

[2] Lynch, J. P., Sundararajan, A., Law, K. H, Kiremidjian, A. S., and Carryer, E., "Power-efficient data management for a wireless structural monitoring system," Proceedings of the 4th International Workshop on Structural Health Monitoring, Stanford, CA, 1177-1184 (2003).

[3] Lynch, J. P. and Loh, K.J. "A summary review of wireless sensors and sensor networks for structural health monitoring". Shock Vibration Digest 38(2), 91-128(2005)

[4] Spencer, B. F., Ruiz-Sandoval, M. and Kurata, N., "Smart sensing technology: Opportunities and challenges". Structural Control and Health Monitoring 11, 349-368(2004)

[5] Candes, E. J., "Compressive sampling," Proceedings of the International Congress of Mathematicians, Madrid, Spain, 1433-1452 (2006).

[6] Donoho, D., "Compressed sensing," IEEE Transactions on Information Theory 52(4), 1289-1306 (2006).

[7] Tropp, J. A. and Gilbert, A. C., "Signal recovery from random measurements via orthogonal matching pursuit, " IEEE Transactions on Information Theory, 53(12), 4655-4666 (2007).

[8] Donoho, D. L. and Tanner, J., "Sparse nonnegative solution of underdetermined linear equations by linear programming," Proceedings of the National Academy of Sciences, 102(27), 9446-9451 (2005).

[9] Chen, S. S., Donoho, D. L., and Saunders, M. A., "Atomic decomposition by basis pursuit, " SIAM Journal on Scientific Computing, 20(1), 33-61(1999)

[10] Ji, S., Xue, Y. and Carin, L., "Bayesian compressive sensing," IEEE Transactions on Signal Processing 56(6), 2346- 2356 (2008)
[11] Tipping, M. E., "Sparse Bayesian learning and the relevance vector machine". Journal of Machine Learning Research 1, 211-244 (2001).
[12] Tipping, M. E., "Bayesian inference: An introduction to principles and practice in machine learning." Advanced Lectures on Machine Learning 3176, 41-62 (2004)
[13] Tipping, M. E. and Faul, A. C., "Fast marginal likelihood maximisation for sparse Bayesian models," Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Key West, FL (2003).
[14] Beck, J. L., "Bayesian system identification based on probability logic", Structural Control and Health Monitoring 17 (7), 825-847(2010).
[15] Ching, J., and Chen, Y.J., "Transitional Markov Chain Monte Carlo method for Bayesian model updating, model class selection and model averaging," Journal of Engineering Mechanics, 133(7), 816-832 (2007).
[16] Beck, J. L., and Au, S.K., "Bayesian updating of structural models and reliability using Markov Chain Monte Carlo simulation," Journal of Engineering Mechanics 128, 380-391 (2002).
[17] Ji, S., Dunson, D., and Carin, L., "Multi-task compressive sensing," IEEE Trans. Signal Process., 57(1), 92-1069(2009).
[18] Bao, Y.Q., Beck, J. L., and Li, H, "Compressive sampling for accelerometer signals in structural health monitoring, " Structural Health Monitoring-An International Journal, 10(3), 235-246 (2011).